Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 5 | May 2017

A SURVEY ON WEB USAGE MINING

Shweta Macwan

Student, Information Technology, Parul University, Vadodara, India - 391760.

ABSTRACT

The extraction of information from the websites and generating relationships from the web data is an important task. The web log file gathers a large amount of information that must be removed using data pre-processing steps. Identifying the importance of website is also crucial. Considering the semantic relationship, hierarchical and non-hierarchical relationships are detected from this log.

KEYWORDS: Web mining, web usage mining, semantic relation, data pre-processing.

INTRODUCTION

Web mining is the use of data mining technique to automatically discover and extract information from the web documents and services. There are two general classes of information that can be discovered by web mining: web activity, from the server logs and web browser activity tracking. Web mining is the use of data mining techniques to automatically discover and extract information from the web documents and services. Good quality data are an important for efficient data analysis. If there is a junk at the input, the same will be at the output, regardless of the method for knowledge extraction used. This applies even more in web log mining, where the log file requires a thorough data preparation. Analyze server access logs and user registration data is also how better to structure a web site for the organization to create a more effective presence can provide information on.

Mining the semantic relations between entities is a vital task in the web mining process. Constructing semantic relationships for structured data from search log takes hierarchical and non-hierarchical relationships. For unstructured data, semantic relationships are created to organize information. Such relations maintain users' perspective. There are many challenges in generating relationships between entities. This is due to explicit, implicit and temporal semantic relations. Considering the temporal relationships, explicit and implicit relationships are created.

In addition to this, the importance of web site has to be generated to increase the quality and business trades of a website in the e-commerce sector.

Data Pre-Processing:

In the real world are incomplete, noisy and inconsistent. The input file for data pre-processing is web server log file of any website. The pre-processing tasks include the following steps:

A. Data Cleaning

This step specifies the filling of missing values; smooth the noisy data identify or remove outliers and solve inconsistency.

B. User Identification

The user identification is done by the user session that is maintained in the user log file. The user is identified by the IP address of the logged in user.

C. Session Identification

The session is identified from the web server log file that stores the information of each session of the user.

D. Path completion

Another important step of the pre-processing step is path completion.

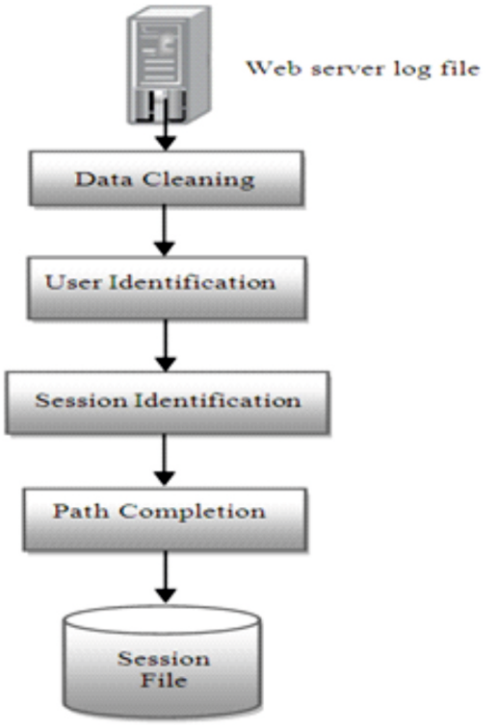

Fig1: Data pre-processing steps

RELATED WORK:

In the recent years, researchers have focused on the analysis of users' behavior using the web mining techniques and methods. Many have tried to combine the techniques of web mining: web content mining, web usage mining and web structure mining to improve the analysis and results. Such tasks helped the researchers to improve the personalization of structure of the website and even the recommending system was improved.

The analysis done through keywords is also an important work done in this field. Hyperlink structure or site structure is also analyzed by using this technique. The problem of back button hyperlink is also solved using the reconstruction of activity approach. Such results are helpful for the commercial benefits of a website and it gives the developer to improve the quality web page but does not reduce the size of the log file.

METHODOLOGY:

The method that is used for improving the quality of a website is k-means clustering algorithm which gave better result than apriori algorithm as well as NMEEF-SD algorithm. IT made five clusters. First cluster was made according to the search engine used where clusters were not obtained. Second cluster collected the session and page views. Third cluster was obtained due to access of keywords. Forth cluster gave lesser instances as compared to the third one. Fifth cluster is similar to second but it is also based on search engine. K-means clustering algorithm proved to better due to its access to the keywords. [1]

For generating temporal semantic relationship, the following method was proposed:

Input: Pair of entities ei and ej and time interval (ts,te)

Output: Semantic context pair including connection entities, the lexical syntactic patterns, context sentences, context graph and context communities from timestamp ts to te. [2]

Data preprocessing step includes data cleaning, user session identification and a new step that is reconstruction of activities of a web visitor for more accurate path completion. This step is used because the backward path is not obtained in the analysis process. The use of back button by users is not stored in log file and hence this approach is used using the hyperlinks from one page to another.

Consider a sequence: A→B→C→D→X. But according to the site map there is no hyperlink from D→X hence this is assumed to be use of back button. Going back, there is no hyperlink from C to X Therefore A→B→C→D →C→X is assumed. Similarly, no hyperlink from B to X. Hence A→B→C→D →C→B→X. Now there is hyperlink from A to X Hence A→B→C→D →C→B→A→X is the final sequence or reconstructed activity of user.[3]

LITERATURE REVIEW:

Table 1: Comparison

Paper	Session Identification	Algorithm used	User Behavior Identification
[1]	Needs to be improved	1) Apriori 2) k-means clustering 3) NMEEF-SD	Not identified
[2]	Not applicable	Temporal semantic relation (TSR)	Not applicable
[3]	Moderate	PageRank algorithm	Identified using reconstruction of users' activity
[4]	Comparatively better	Simulated Annealing	Identified using web sessionization
[5]	Not applicable	1) Semantic Content Relationship (SCR) 2) Query Log Graphs (QLG)	Not applicable

Table 2

Paper	Accuracy	Computational time
[1]	Better using k-means clustering algorithm	Not applicable
[2]	High	Moderate
[3]	Moderate	High
[4]	Not achieved	High
[5]	Higher using QLG	Better using SCR

CONCLUSION:

The increasing popularity of the Web has greatly attracted the Web mining technology. A vital research area in Web mining is Web usage mining which mainly focuses on the discovery of patterns in the browsing and navigation data of Web users. Web usage mining has been a potential technology for understanding behavior of the user on the web. There are several techniques proposed by the researchers for the web usage mining that improves the quality and design of the web page also gives semantic relationships between entities. The pre-processing, pattern preprocessing and pattern analysis are important steps in the web usage mining. The pre-processing of data gives the accuracy in analysis of the website. It is obvious that enhanced cluster recovery provides highly accurate guessing of a web user's future visit if the user's cluster can be exactly determined. As a future work, integrating self-organizing approach can improve the quality of the analysis. According to accuracy of reconstruction of session, processing time can be increased. Keywords can be improved by concentrating on the visits obtained by the other websites.

REFERENCES:

[1] C.J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M.J. del Jesus, S. García, "Web usage mining to improve the design of an e-commerce website: orolivesur.com", EXPERT SYSTEMS WITH APPLICATIONS, VOLUME 39, ISSUE 12, 15 SEPTEMBER 2012

[2] Zheng Xu, Xiangfeng Luo, Shunxiang Zhang, Xiao Wei, Lin Mei, Chuanping Hu, "Mining temporal explicit and implicit semantic relations between entities using web search engines", FUTURE GENERATION COMPUTER SYSTEMS, VOLUME 37, JULY 2014

[3] Kapusta Jozef, Munk Michal, Drlík Martin, "Identification of Underestimated and Overestimated Web Pages Using pagerank and Web Usage Mining Methods", TRANSACTIONS ON COMPUTATIONAL COLLECTIVE INTELLIGENCE XVIII, LECTURE NOTES IN COMPUTER SCIENCE, SPRINGER BERLIN HEIDELBERG, VOLUME 9240, JULY 2015

[4] Tomás Arce, Pablo E. Román, Juan Velásquez, Víctor Parada, "Identifying web sessions with simulated annealing", EXPERT SYSTEMS WITH APPLICATIONS, Volume 41, Issue 4, Part 2, March 2014

[5] Pei-Ling Hsu, Hsiao-Shan Hsieh, Jheng-He Liang, Yi-Shin Chen, "Mining various semantic relationships from unstructured user-generated web data", WEB SEMANTICS: SCIENCE, SERVICES AND AGENTS ON THE WORLD WIDE WEB, Volume 31, March 2015